

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 22 Number 12, December 2017

ISSN 1531-7714

Developing Rubrics to Assess Complex (Generic) Skills in the Classroom: How to Distinguish Skills' Mastery Levels?

Ellen Rusman, *Welten Institute, Open University of the Netherlands*

Kim Dirkx, *Welten Institute, Open University of the Netherlands*

Many schools use analytic rubrics to (formatively) assess complex, generic or transversal (21st century) skills, such as collaborating and presenting. In rubrics, performance indicators on different levels of mastering a skill (e.g., novice, practiced, advanced, talented) are described. However, the dimensions used to describe the different mastery levels vary within and across rubrics and are in many cases not consistent, concise and often trivial, thereby hampering the quality of rubrics used to learn and assess complex skills. In this study we reviewed 600 rubrics available in three international databases (Rubistar, For All Rubrics, i-rubrics) and analyzed the dimensions found within 12 strictly selected rubrics that are currently used to distinguish mastery levels and describe performance indicators for the skill 'collaboration' at secondary schools. These dimensions were subsequently defined and categorized. This resulted in 13 different dimensions, clustered in 6 categories, feasible for defining skills' mastery levels in rubrics. The identified dimensions can specifically support both teachers and researchers to construct, review and investigate performance indicators for each mastery level of a complex skill. On a more general level, they can support analysis of the overall quality of analytic rubrics to (formatively) assess complex skills.

Complex generic skills, such as collaborating, presenting and information literacy, are gaining increased attention in education. These skills, although not new in their nature and disposition, are at national and international policy level increasingly seen as important and labeled as '21st century skills'. 21st century skills are transversal or generic complex skills (and associated knowledge and attitudes) which are generally seen as necessary in order to live, work and contribute to the current and future knowledge society (Voogt & Pareja Roblin, 2012). These generic skills are not specific for a domain, occupation or type of task, but important for all kinds of work, education, and life in general (Bowman, 2010) and are applicable in a broad range of situations and subject domains. They consist of constituent sub-skills which concertation require high cognitive effort, concentration (Galligan, Maskery, Spence, Howe, Barry, Ruston & Crawford,

2000; Kirschner & van Merriënboer, 2008) and prolonged repetitive practice in order to master them.

Although the importance of learning generic complex skills is evident, many schools are struggling with the question how to address them in their curriculum (Rusman, Martínez-Monés, Boon, Rodríguez-Triana & Villagrà-Sobrino, 2014). In many schools, rubrics are used as an instrument to support learning complex skills in schools. Although rubrics are a suitable way to support skill learning, their effect and feasibility largely depends on the quality of the used rubrics. In this study we review the quality of rubrics which are developed, used and shared by educators from various countries in three large international databases: Rubistar, For All Rubrics and i-rubrics. These databases contained mainly rubrics in English. We focused on the consistency of performance indicators used to define the different levels of mastery

of a skill, one of the main components of rubrics' quality, and assess the dimensions used in order to describe them. We address this topic as the design and use of consistent and continuous dimensions underlying the description of performance indicators for each mastery level of a skill is a re-occurring problem in rubric design (Harper, O'Connor, & Simpson, 1999; Wiggins; 1998) and there are almost no existing guidelines available to support their design. Based on the review of 'verbal qualifiers' (Rohrman, 2002) which are used in and across the descriptions of the different skills mastery levels in a rubric, we subsequently derived various dimensions that can be used to consistently define performance indicators across the mastery level descriptions of a skill. These dimensions can be used both for rubric design as well as for the evaluation of the quality of mastery level descriptions within a rubric.

Rubrics for Learning Complex Skills

In order to learn complex generic skills, like for example collaboration, learners need to acquire a concrete and consistent mental model of what it means to master this skill. Additionally, when learners know and understand the gap between their current and targeted mastery level, they can better determine, eventually with the help of the teacher, their subsequent learning activities. Rubrics are a valuable

instrument to support learners while developing and (formatively) assessing complex skills (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013) as they support both mental model formation as well as feedback and self-regulation processes, e.g. by means of self-, peer and expert feedback and reflection on performances. Although both 'analytic' and 'holistic' rubrics are distinguished, in this article we focus on the analytic variant, as these are the most commonly used rubrics. Analytic rubrics are more explicitly and precisely describing performance criteria, which in turn support mental model formation (Arter & Chappuis, 2006).

An analytic rubric (see Figure 1 for an example) provides textual descriptions of skills' mastery levels with performance indicators that describe concrete behaviour for all constituent sub-skills at each mastery level (Reddy, 2011; Sluijsmans, Joosten-ten Brinke & van der Vleuten, 2013). Such performance indicators specify aspects of variation in the complexity of a sub-skill (e.g., presenting for a small, homogeneous group compared to a more complex presentation for a large heterogeneous group) and related mastery levels (Kirschner & van Merriënboer, 2008) in concrete terms. The performance indicators are the criteria that an assessor considers when determining the quality of a pupils' work.

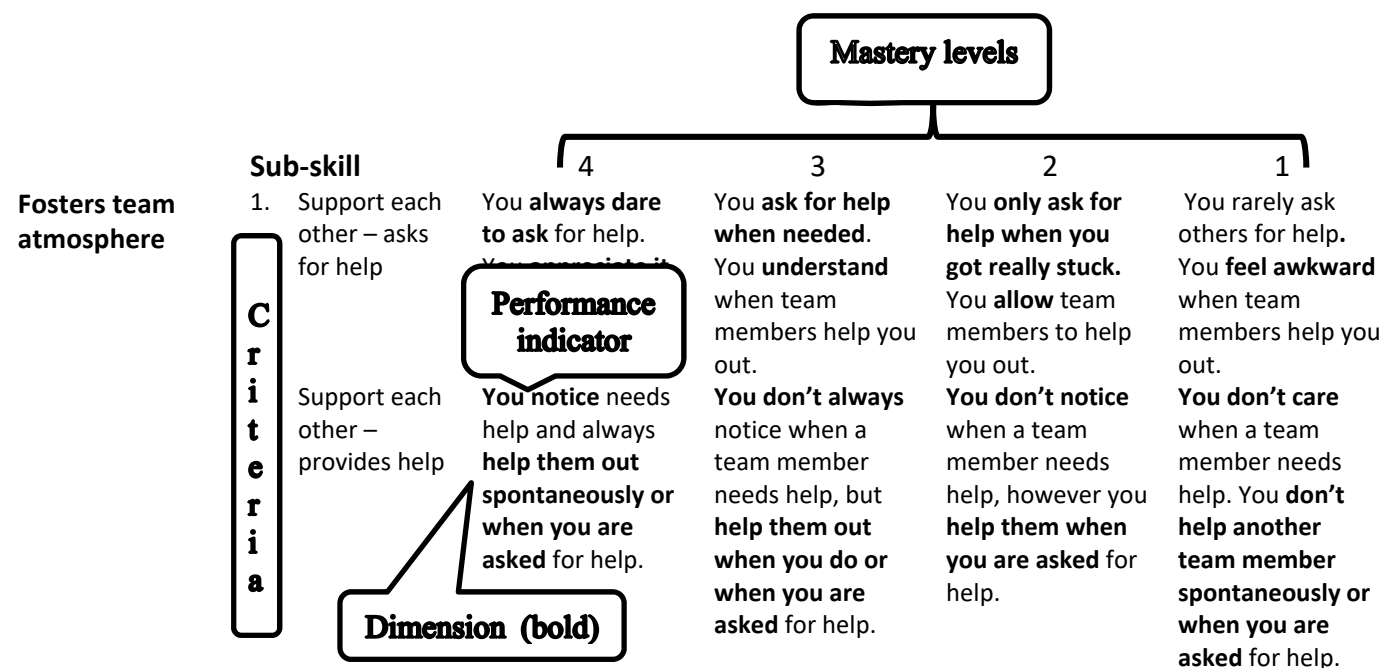


Figure 1. Example rubric to clarify terminology

Performance indicators make the norm for a skills' mastery level explicit for pupils (Jonsson & Svingby, 2007) and help learners to gain a clear(er) picture of the strived-for behavior, so that learning objectives are clear and transparent while practicing a skill (Jonsson & Svingby, 2007; Reddy & Andrade, 2010). Moreover, the usage of analytic rubrics can support various learning and assessment processes relevant when learning a complex skill, such as self-regulated learning, (self-, peer-, and expert-) feedback, and reflection on performances. Review studies (Panadero & Jonsson, 2013; Schildkamp, Heitink, van Der Kleij, Hoogland, Dijkstra & Kippers, 2014) show for example that rubrics are especially suitable for communication between relevant actors (e.g., learners (amongst each other), teachers, experts) about the requirements and expectations of a skill's mastery levels and in this way contribute to the regulation and adjustment of the individual learning process. This enables learners, while practicing a skill, to pay extra attention to the aspects of a skill they didn't master very well yet and to self-regulate their learning process, as they use the rubric to mirror and provide guidance to define the learning objectives and steps they still need to take in order to reach the desired mastery level of a skill. Finally, an analytic rubric facilitates a more transparent and reliable assessment (Kerkhoffs, Starks & Zeelenberg, 2006).

Designing Rubrics and the Problem of Articulating Mastery Level Descriptions

Since rubrics are multi-faceted instruments and can be used to enhance learning and assessment practices in various manners (e.g. for self-, peer- and expert-assessment), there is an expanded use of analytic rubrics in education. Teachers often re-use existing rubrics, but also create their own. However, although rubrics are seen as very helpful tools for assessment, they only lead to more transparent and reliable evaluation of performances when they are themselves valid and reliable and have clear, consistent, concise/focused and usable criteria (Brookhart & Chen, 2015). Therefore, it is important to consider various quality criteria carefully when designing or evaluating rubrics. This is for example done by Arter and Chappuis (2006) who developed a framework with several quality criteria to evaluate the overall quality of rubrics. They came up with two main criteria: coverage and clarity, which they further specified and broke down. Coverage, referring to whether a rubric covers the 'right' content, for example is specified by

'organization of criteria' and 'number of levels fits target and usage', whereas 'clarity' refers to whether the rubric clearly conveys to users what it is meant to, also referring to 'definition' and 'parallelism' of rubrics' mastery level descriptions.

These more generic quality issues of rubric design are also addressed by other authors (Arter & McTighe, 2001; Harper et al., 1999; Moskal, 2003; Popham, 1997; Stiggins, 2001; Wiggins, 1998). They, for example, suggest that indicators should be clearly formulated, that language used should relate to the target group since it affects the reliability of the assessment made by both pupils and raters (Arter & McTighe, 2001; Arter & Chappuis, 2006; Harper et al., 1999; Moskal, 2003; Popham, 1997; Stiggins, 2001; Wiggins, 1998). Others note that a balance between generalized wording, which increases usability, and detailed descriptions, which ensure greater reliability, must be achieved (Popham, 1997; Simon & Forgette-Giroux, 2001; Wiggins, 1998).

Several authors also repetitively stress that the mastery levels, the various levels of accomplishment, should be clearly and meaningfully differentiated and based on an unidimensional construct (Gerbing & Anderson, 1988; Moskal, 2003; Wiggins, 1998). A 'dimension' (see the 'bold' words in Table 1) is the translation of this underlying construct in two or more measures (or alternative indicators) for each mastery level, which provide an empirical estimate of the theoretical construct of interest (Gerbing & Anderson, 1988).

Another important issue that Arter and Chappuis (2006) address is the parallelism of the performance level descriptions. They emphasize that the levels in a rubric should be parallel in content, which means that if an indicator of quality is discussed in one level, it is discussed in all levels. However, in many rubrics used in daily teaching practices there is a great inconsistency in the descriptions of performance criteria across skills' levels. There are no standardized, universally understood dimensions available that can be used to describe performance indicators for each mastery level of a constituent subskill within rubrics (Tierney & Simon, 2004). In 1997, Popham already warned that rubrics can harm learning when there are a lot of inconsistencies in the performance indicator descriptions across mastery levels. There is an urge for consistent and concise wording to describe performances across the levels of achievement (Harper et al., 1999; Wiggins; 1998). It is important that the

attribute, or underlying characteristic of each performance criterion is explicitly chosen and designed within the rubric and that the words that describe the changing values of the attribute (e.g., few, most, all) are differentiations along a consistent dimension. Simon & Forgette-Giroux (2001) suggest that the descriptors for each level should deal with the same performance criteria and attributes in order for the progressive dimensions to be continuous and consistent from one level to the other. One of the major causes for problems in rubric design is that there is very little research on how to formulate clear, meaningful, unidimensional and differentiating dimensions that are used to describe skills' mastery levels (Reddy, 2011; Tierney & Simon, 2004). And although these problems were identified, no specific guidelines to overcome these problems are formulated yet. By reviewing and defining dimensions that are currently used and that can be used to consistently define performance indicators across the mastery levels of a constituent sub-skill, we aim to improve the rubric design process.

Two of the few sources that also address this design issue are those of Aiken (1996) and Rohrmann (2002). They refer to three dimensions for the design of performance indicators that are commonly used. These are: 1) amount, 2) frequency, and 3) intensity. Tierney and Simon (2004) linked these scales to different attributes. These attributes are breath (linked to amount or quantity), accuracy (linked to frequency), relevance and clarity (linked to intensity). However, the studies of Aiken (1996) and Rohrmann (2002) focused on the construction of rating scales (in the context of questionnaires) and the differentiating criteria found were mainly quantitative. Various researchers investigating skill development and rubric design on the contrary stress that the performance indicators of mastery levels should be described qualitatively (e.g., Arter & Chappuis, 2006). For example, Bulthuis (2013) and Van Merriënboer, Clark and Croocke (2002) propose that the degree of independency of a pupil and task complexity should play an important role when describing mastery levels of the constituent subskills of a skill.

With this study, we aim to contribute to the design of high quality rubrics by providing insight in the range and types of dimensions used to formulate performance indicators and their frequency of use in practice. We focused on rubrics for 'collaboration' and

'teamwork' and where guided by the following research questions:

1. What dimensions can be detected in existing rubrics to differentiate between the various mastery levels of a skill?
2. What are the most frequently used dimensions in existing rubrics to differentiate between the various mastery levels of a skill?

Methodology

In order to answer the research questions, we searched for rubrics focusing on the assessment of collaboration skills in three different international databases (i.e., Rubistar, For All Rubrics, i-rubrics). We used the following search criteria (collaborate, collaboration, collaborative working skills, group work) and received in October 2016 more than 50.000 hits across these three databases. Therefore, we needed to narrow down our search (see Table 1) and did so by using very strict selection criteria, in order to select the rubrics with the highest quality. As a first step we looked at the different search results and noticed a lot of overlap between the rubrics that were found (e.g., same rubrics registered, sometimes with small changes, by different authors). Based on this information we agreed to use only the first 200 results of each database and for each search term. Furthermore, we only included rubrics that were published within the last five years. This filtering led to 522 results. Then, we excluded rubrics that had no unique titles and descriptions which resulted in 133 rubrics. To further narrow down our search, we used additional selection criteria. For inclusion in our study, a rubric should:

- be aimed at higher secondary education (1st two classes)
- be generic instead of relating to domain-specific collaboration
- be aimed at collaboration skills instead of collaborative learning
- address the collaboration processes, not the product that results from collaboration
- look at individual performance in the group (not at the group as a whole)
- address the complete construct of collaboration
- distinguish several constituent sub-skills

- describe sub-skills and mastery levels in terms of concrete behaviour in an analytic format (not holistic)
- be publicly accessible for teachers
- be unique (i.e., not based on already included rubrics)
- be suited for summative and formative assessment methods (e.g., self-, peer-, and teacher-based assessment)
 - fulfill several quality criteria (Arter & Chappuis, 2006): consistency between performance criteria
 - independency of specified criteria and scales
 - transparency
 - parallelism

Table 1 shows how the application of these criteria narrowed down the number of rubrics used within this study. Based on these strict criteria, 12 rubrics (English) which fulfilled all generic and quality criteria were left. As an existing schema for analyzing the results was not available (i.e., Aiken (1996) and Rohrmann (2002) only defined quantitative scales), we used a bottom-up

procedure to analyze the rubrics which is comparable to the one used by Tierney & Simon (2004). First we looked at the verbal qualifiers (Rohrmann, 2002) that were used in these 12 rubrics to describe a certain mastery level. Each of the two involved researchers analyzed the rubric individually. Then, the rubrics and the verbal qualifiers that were found were discussed during a two-hour meeting in order to find agreement on the dimensions that were used to describe the performance indicators of each mastery level in the rubric. Words such as “occasionally” or “some of the time” were clustered as belonging to one dimension. Then, a preliminary label was given to the verbal qualifiers based on the function of a dimension. “Occasionally” and “some of the time” refer for example to the ‘regularity’ by which the behavior is shown, whereas “five time” or “two times” refer to the ‘frequency’ by which certain behavior is shown. Then, the other rubrics were scored according to the agreed dimensions and examples of every dimension were included. In three cases, new dimensions were found in the rubrics. These were also listed. Subsequently, all dimensions were discussed to settle on them and to provide them with a name and a description. Finally, one of the two involved researchers counted how often the dimensions were used in the selected set of rubrics.

Table 1. Overview of the rubric selection process

Search term	Initially found (>searched in first 200 results)	Selection based on title	Final selection based on face validity
Collaborate rubistar	6	1	-
Collaborate four all rubric	2	2	-
Collaborate i-rubric**	-	-	-
Collaboration rubistar	63	11	1
Collaboration four all rubric	2	2	2
Collaboration i-rubric**	2046>200	-	-
Collaborative working skills rubistar	74	22	2
Collaborative work skills four all rubric	1	1	1
Collaborative work skills i-rubric**	41852>200	-	-
Group work rubistar	5491-> 200	9	5
Group work four all rubric	4	2	1
Group work i-rubric**	42144	-	-
Total	552	133	12

** i-rubric contained many similar rubrics as rubistar and did not lead to finding new rubrics.

Results

We found 13 unique dimensions in the example rubrics and clustered them into six categories: measurement (4 dimensions), independence (2 dimensions), quality aspects (2 dimensions), values of the pupil (1 dimension), engagement (2 dimensions)

and complexity (2 dimensions). The categories and dimensions are displayed in Table 2 with a short description and some example ‘verbal qualifiers’ on which we based our categorization.

The results show that there are quite some unique dimensions to indicate differences between pupils’

Table 2. Overview of dimension categories and underlying dimensions

Category (Count)	Dimension	Description	Examples (verbal qualifiers)
Measurement (60)	Regularity	A measure that expresses to what extent the behavior is present all of the time.	occasionally; sometimes; rarely; often; almost always; never; usually; some of the time
	Frequency	A measure that expresses how often the behavior is shown.	five times, two times.
	Quantity	A scalable measure that expresses the degree of the behavior.	little, none; small mistakes; minor mistakes; 90-100%; majority, half; each idea-few ideas
	Consistency	A measure that expresses the degree to which the behavior is automated as indicated by the degree to which a pupil shows the same behavior in different situations.	some of the time; consistently; routinely; most of the time; marginally; mostly.
Independence (12)	Guidance	An adjective that indicates the degree to which a pupil needs hints or clues from others to perform a task.	when prompted; requires guidance; needs to be reminded to stay on task.
	Support	An adjective that indicates the degree to which a pupil needs concrete input of others to perform a task.	others check or re-do a task.
Quality aspects(28)	Quality of behavior	A description of the degree in which the behavior fulfills the conceptualization of professional behavior.	asks challenging questions; asks questions to elicit new information; listens respectfully and modifies own thinking; is able to accurately restate the alternative viewpoint on issue; provides work of the highest quality; wastes time-uses time well; respects other perspectives; helps the group; assumes various roles; often not a good team member.
	Perceived usefulness of the behavior	An adjective that refers to the effectivity of the behavior in relation to the intended aim.	effective; productive; useful; feasible; specific; supportive; appropriate; accurate.
Values of a person (2)	Internalization of professional behavior/ Attitude	A description that expresses the degree to which a pupil shows the aimed behavior because of his/her own convictions.	values collaboration; genuine effort; honestly.
Engagement (28)	Pro-activity	A description that expresses the degree to which a pupil contributes to the group or shows aimed behavior out of his own.	actively works; unless required; contributes a lot of effort; refuses to participate; actively looks for; refines; try out solutions suggested by others; others can count on; group members must nag, prod and remind to stay on task; is reluctant; unless required.
	Involvement	An adjective that indicates the degree to which a pupil is willing to put effort in an activity.	best effort-strong effort-some effort -little effort.
Complexity (4)	Detailedness	A description of the degree to which a pupil takes into account the depth or width of the situation.	detail; in depth; not in detail.
	Integration	A description that indicates the degree to which a pupil brings the pieces in a project or task together.	separately; puts together.

mastery levels. Whereas measurement issues refer to rather quantitative differences (see also Aiken, 1996; Rohrmann, 2002), the other clusters refer to more qualitative differences between pupils, such as the quality of their behavior, the level of engagement a pupil shows when working on a task with others or how much support a pupil needs. Especially these qualitative indicators are important to distinguish high performing pupils from low performing pupils (e.g., Brookhart, 1999). Moreover, the degree to which the pupil has internalized the targeted behavior seems to be important in rubrics on collaborative skills as evidenced by the dimensions ‘values of a pupil’ and ‘engagement’. It seems to be not enough that a pupil is only performing conform the standards, but the behavior should become part of pupils’ regular behavioral pattern. This is in line with the conceptualization of competences (Bulthuis, 2013; van Merriënboer et al., 2002).

The consistency between the dimensions used within the performance indicator descriptions within and across rubrics was quite poor. Most rubrics used words signaling a mastery level only in one or two performance indicators per constituent sub-skills. Also across rubrics, many different verbal qualifiers or signaling words were used to describe the same scales.

Conclusion and Discussion

In previous research on rubric development, there has been given little attention to the design of performance indicators of skills’ mastery levels and underlying dimensions used to formulate descriptions. However, it is important that there is consistency in the usage and verbal qualifiers of such dimensions (e.g., Tierney & Simon, 2004) in and between rubrics. In the present research we investigated which dimensions could be detected in existing rubrics for the skill ‘collaboration’. Although the consistency of the use of the dimensions to describe performance indicators associated with mastery levels was low in the analyzed rubrics, we found a broad range of qualitative and quantitative dimensions. We also found a tendency to use quantitative dimensions to define performance indicators, referring mainly to measurable features (e.g., frequency). These dimensions overlap somewhat with the ones found by Aiken (1996) (and Rohrmann (2002)). However, research on rubric design suggests the use of more qualitative dimensions to differentiate between mastery levels (e.g., Arter & Chappuis, 2006;

Brookhart, 1999), such as ‘Independency’ and ‘Task complexity’ (Bulthuis, 2013). These qualitative dimensions describe standards of consistent, concrete, visible behavior of mastering a skill to mirror and compare pupils performances against. These qualitative attributes, like independency, also reflect principles of the 4C/ID model (van Merriënboer et al., 2002). This model claims that, in order to acquire a certain complex skill, learning tasks should decline over time with regard to the level of guidance and increase in the complexity of tasks offered to learners. Accordingly, learners should practice tasks of the same complexity level until they can perform them without guidance and then proceed to a more complex task. Important hereby is the variability of the learning tasks, meaning that learners should practice with tasks that represent different conditions encountered in professional practice. Whereas attributes referring to guidance or support are clearly seen in the analyzed rubrics, attributes that refer to task complexity are less clearly seen in existing rubrics. In the existing rubrics, the focus is more on the behavior of the student (e.g., does he/she bring the pieces in a project or task together) instead of characteristics of the context (how complex is the context in which the task should be performed, e.g. in terms of number of people involved, number of possible solutions for the problem).

Looking at the theory of competency development (Bulthuis 2013; van Merriënboer et al., 2002), other potentially relevant and additional attributes are the degree of automatization of routine or automated (sub-) skills and the degree of integration, ‘concertation’ of all sub-skills.

For further development and research on rubric design, it would be worthwhile to approach scale development of rubrics from a more theoretical perspective, as done in an explorative way above. Another line for future research would be to see whether more and other dimensions could be identified when looking at rubrics for other complex skills (e.g. presenting, problem solving, information literacy) instead of those found for ‘collaboration’, in order to provide uniform and usable guidelines for the design of rubrics and performance indicators’ underlying dimensions.

References

- Aiken, L. R. (1996). *Rating scales and checklists: Evaluating behavior, personality, and attitudes*. New York, NY: John Wiley & Sons, Inc.
- Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving pupil performance*. Thousand Oaks, CA: Corwin Press/Sage Publications.
- Arter, J. A., & Chappuis, J. (2006). *Creating & recognizing quality rubrics*. Boston: Pearson.
- Brookhart, S. M. (1999). The Art and Science of Classroom Assessment: The Missing Part of Pedagogy. ASHE-ERIC report. Higher Education Report (Vol. 27, No.1). Washington, DC: The George Washington University, Graduate School of Education and Human Development. Retrieved on 5 September 2016 from <http://files.eric.ed.gov/fulltext/ED432937.pdf>
- Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review* 67, 343-368. <http://dx.doi.org/10.1080/00131911.2014.929565>
- Bowman, K. (2010). Background paper for the AQF Council on generic skills. Retrieved on 24 October 2016 from <http://www.aqf.edu.au/wp-content/uploads/2013/06/Generic-skills-background-paper-FINAL.pdf>
- Bulthuis, P. (2013). Het ZelCommodel, grip op competentieniveaus [The selfcommodel, defining competence levels]. *Examens*, 2, 5-10. Retrieved on 3 November 2016 from <https://score.hva.nl/Bronnen/Het%20ZelCommodel%20-%20Grip%20op%20competentieniveaus.pdf>
- Galligan, D., Maskery, F., Spence, C., Howe, J., Barry, D., Ruston, T., & Crawford, A. (2000). *Advanced PE for Edexcel*. Oxford: Heinemann Educational Publishers.
- Goodrich, H. (1997). Understanding rubrics. *Educational Leadership*, 14-17. Retrieved on 24 October 2016 from http://tccl.rit.albany.edu/knilt/images/6/63/Andrade_Rubrics.pdf
- Gerbing, D. W., & Anderson, J. C. (1988). An Updated Paradigm for Scale Development Incorporating Unidimensionality and its assessment. *Journal of Marketing Research (JMR)*, 25 (2), 186-192. Retrieved on 2 December 2016 <http://rhowell.ba.ttu.edu/gerbinganderson.pdf>
- Harper, M., O'Connor, K. & Simpson, M. (1999). *Quality assessment: Fitting the pieces together*. Toronto, ON: Ontario Secondary School Teachers Federation.
- Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144. doi: 10.1016/j.edurev.2007.05.002
- Kerkhoffs, J., Starks, E. & Zeelenberg, T. (2006). Rubrics als beoordelingsinstrument voor vaardigheden. Enschede, Nederland: SLO. Retrieved on 17 October 2016 from <http://www.slo.nl/downloads/rubricsals.pdf>
- Kirschner, P. A., & van Merriënboer, J. (2008). "Ten Steps to Complex Learning A New Approach to Instruction and Instructional Design". In T. L. Good (Ed.), *21st century education: A reference handbook* (pp. 244–253). Thousand Oaks, CA: Sage. Retrieved on 23 September 2016 from <http://web.mit.edu/xtalks/TenStepsToComplexLearning-Kirschner-VanMerrienboer.pdf>
- Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation*, 8, (14). Retrieved on 3 December 2016 from <http://pareonline.net/getvn.asp?v=8&n=14>
- Panadero, E. & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144. doi: 10.1016/j.edurev.2013.01.002
- Popham, W. J. (1997). What's wrong - and what's right - with rubrics. *Educational Leadership*, 55, 72-75. Retrieved on 4 December 2016 from <http://pareonline.net/getvn.asp?v=9&n=2>
- Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment Evaluation in Higher Education*, 35(4), 435-448. doi: 10.1080/02602930902862859
- Reddy, Y.M. (2011). Design and development of rubrics to improve assessment outcomes. *Quality Assurance in Education*, 19(1), 84 – 104. <http://dx.doi.org/10.1108/096848811111107771>
- Rohrmann, B. (2002). Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data. Retrieved 7 December 2016 from <https://usm.maine.edu/sites/default/files/assessment/Rubric%20Article%20Tierney-Simon2004.pdf>
- Rusman, E., Martínez-monés, A., Boon, J., Rodríguez-triana, M. J. & Villagrà-sobrino, S. (2014). Gauging Teachers' Needs with Regard to Technology- Enhanced Formative Assessment (TEFA) of 21st Century Skills in the Classroom. In: Computer Assisted Assessment. Research into E-Assessment. Proceedings of the International Computer Assisted Assessment (CAA)

- Conference (pp. 1–14). Retrieved on 18 August 2016 from
https://www.researchgate.net/profile/Maria_Rodriguez-Triana/publication/289710915_Gauging_Teachers'_Needs_with_Regard_to_Technology-Enhanced_Formative_Assessment_TEFA_of_21st_Century_Skills_in_the_Classroom/links/56e6c58608ae65dd4cc1cde2.pdf?origin=publication_list
- Schildkamp, K., Heitink, M., Van Der Kleij, F., Hoogland, I., Dijkstra, A., & Kippers, W. (2014). Voorwaarden voor effectieve toetsing. Een praktische review. [Prerequisites for effective testing. A practical review]. Den Haag: Nationaal Regieorgaan Onderwijsonderzoek. Retrieved on February 2016 from
<http://doc.utwente.nl/94424/1/Eindrapport-formatief-toetsen-Utwente-revisie.pdf>
- Simon, M. & Forgette-Giroux, R. (2001). A rubric for scoring postsecondary academic skills. *Practical Assessment, Research & Evaluation*, 7(18). Retrieved on 2 December 2016 from
<http://pareonline.net/getvn.asp?v=7&n=18>.
- Sluijsmans, D., Joosten-ten Brinke, D., & Van der Vleuten, C. (2013). Toetsen met leerwaarde [Assessments with value for learning]. Retrieved on 24 October 2016 from
<https://www.nro.nl/wpcontent/uploads/2014/05/PR OO+Toetsen+met+leerwaarde+Dominique+Sluijsmans+ea.pdf>
- Stiggins, R. J. (2001). *Student-involved classroom assessment (3rd ed.)*. Upper Saddle River, New York: Prentice-Hall, Inc.
- Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research and Evaluation*, 9 (2). Retrieved on 15th of November 2016 from
<http://PAREonline.net/getvn.asp?v=149&n=142>
- van Merriënboer, J. J. G., Clark, R. E., & de Croock, M. B. M. (2002). Blueprints for complex learning: The 4C/ID-Model. *Educational Technology, Research and Development*, 50(2), 39-64. Retrieved on 20 January 2016 from
http://www.cogtech.usc.edu/publications/clark_4cid.pdf
- Voogt, J., & Pareja Roblin, N. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, 4(3), 299-321.
<http://dx.doi.org/10.1080/00220272.2012.668938>
- Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve pupil performance*. San Francisco, CA: Jossey-Bass Publishers.
- Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing pupil learning. *The Journal of Effective Teaching*, 7(1), 3-14. Retrieved on 5 January 2017 from
http://www.uncw.edu/jet/articles/vol7_1/wolf.pdf

Citation:

Rusman, Ellen & Dirkx, Kim. (2017). Developing Rubrics to Assess Complex (Generic) Skills in the Classroom: How to Distinguish Skills' Mastery Levels?. *Practical Assessment, Research & Evaluation*, 22(12). Available online: <http://pareonline.net/getvn.asp?v=22&n=12>

Corresponding Author

Ellen Rusman
Welten Institute - Open University of the Netherlands
Postbus 2960 6401 DL Heerlen
email: ellen.rusman [at] ou.nl